



工业互联网产业联盟
Alliance of Industrial Internet

Industrial Big Data: Technology and Application

工业大数据：技术与应用

Chen Wang 王晨

National Engineering Lab for Big Data Software,
Tsinghua University

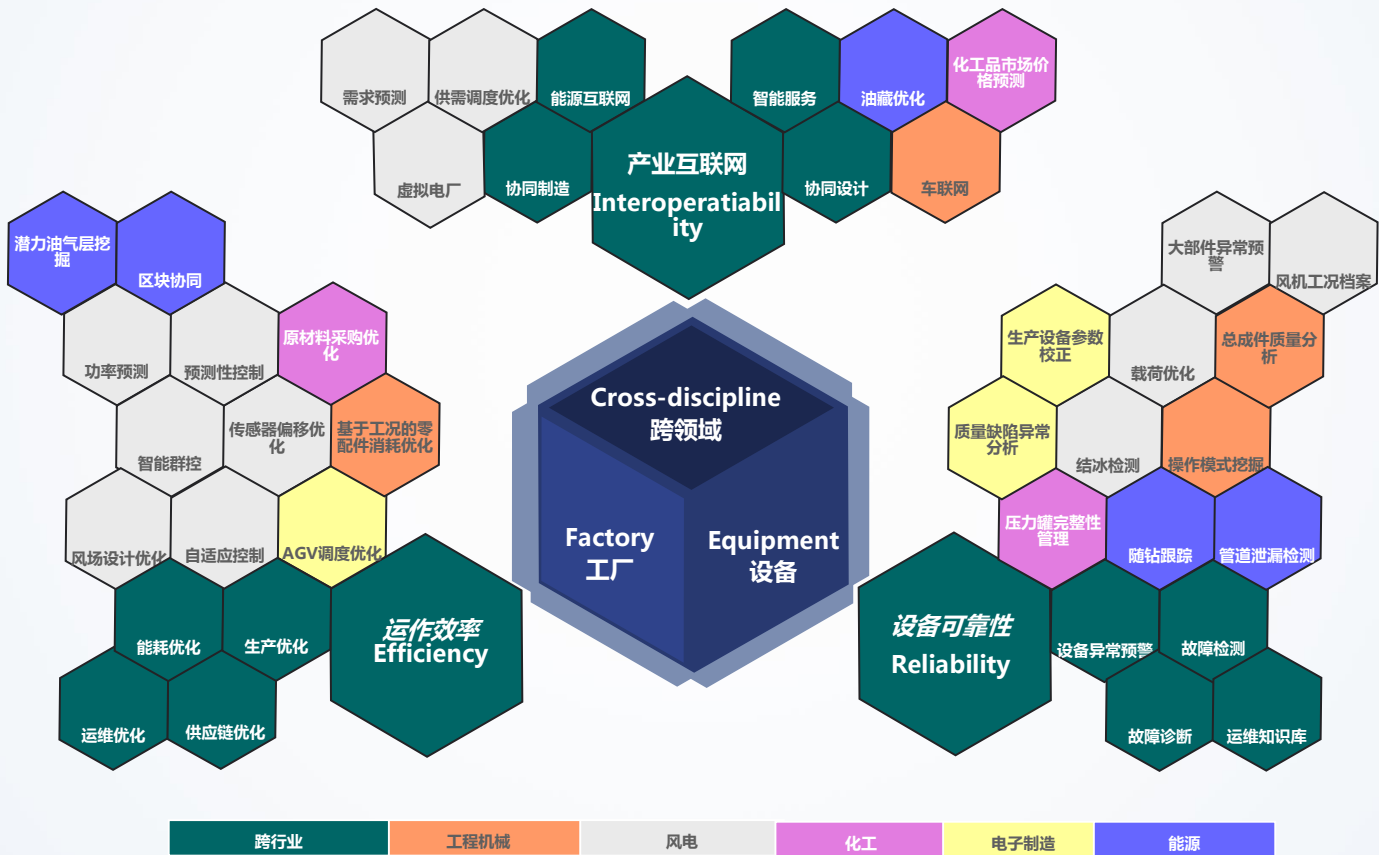
清华大学大数据系统软件国家工程实验室

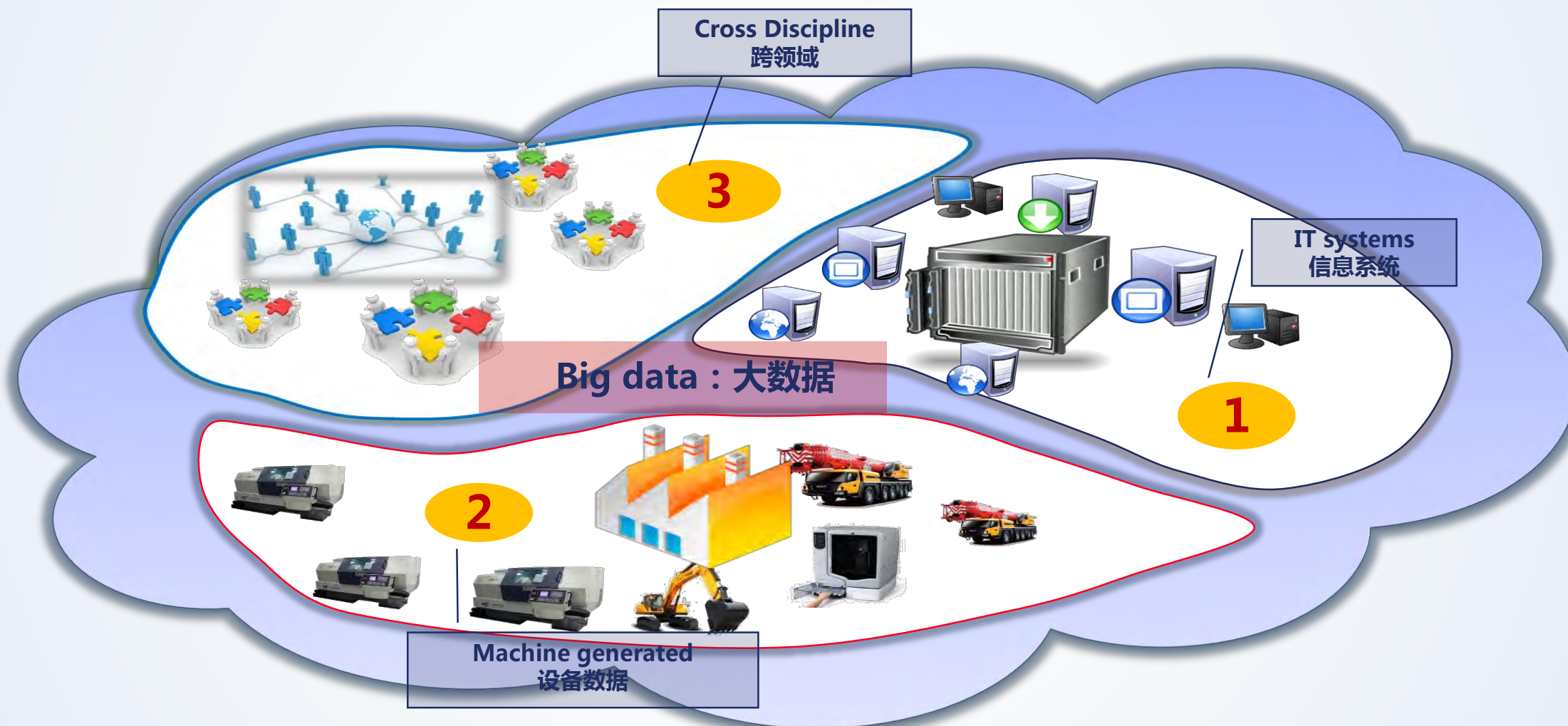
Innovation Institute for Industrial Big Data

工业大数据创新中心

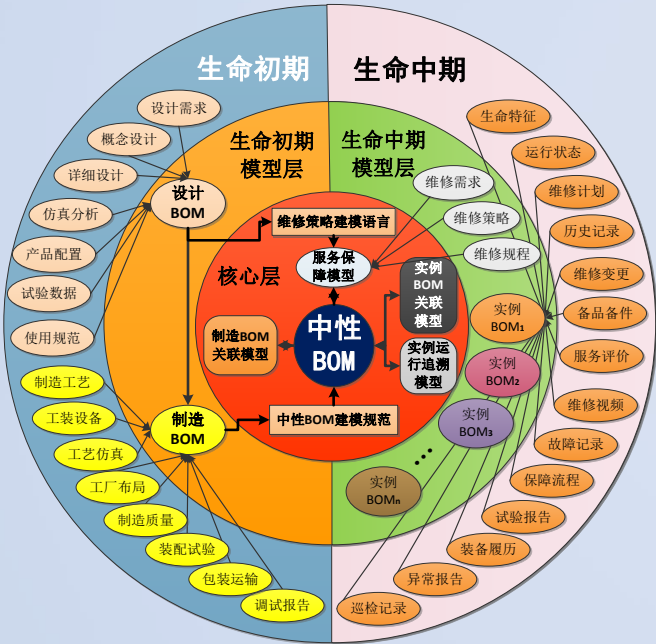
Application Areas of Industrial Big Data

工业大数据应用领域





以多种类非结构化工程数据、过程与BOM图数据、高端装备监测时序数据为代表的高端制造业大数据呈现“多模态(Multimodal)，高通量(High throughput)，强关联(Strong Correlation)”特性



数据模态多样,结构关系复杂
典型高端制造企业数据类型可达300余种
汽轮机35万个零部件数据

海量高速
机器7*24产生
采集频率高，数据量大



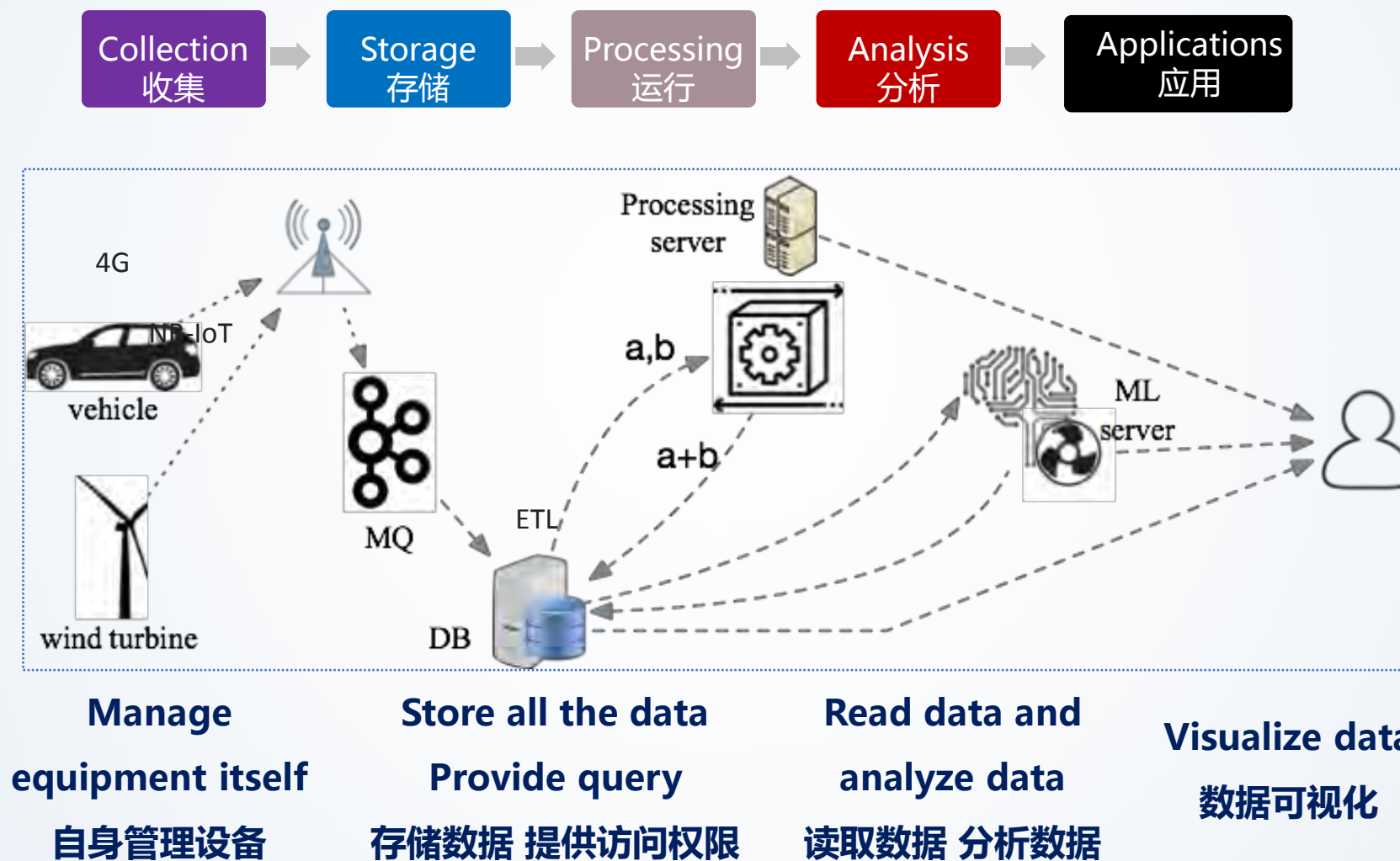
数据通量大
50Hz，500测点/台，2万台风机
最高可达数千万数据点/秒



多学科异构数据信息交互模型

协作专业多
飞行器研发相关专业200多类

Life Cycle of Time Series Data 时间序列数据的生命周期



Huge amount of Time Series data

大量的时间序列数据

- Lots of devices 许多的设备
- Many sensors on a device 一个设备的多个传感器
- Generate data in a high frequency 高频率生成数据
- For monitoring, forewarning, and improvement 为了显示, 预警, 升级等目的



- About 120~510 variables/sensors per wind turbine 大约 120~510变量/传感器每个风力涡轮机
- Frequency ranging from 0.00167 Hz to 50Hz 频率范围从 0.00167赫兹到50赫兹
- More than 20,000 wind turbines 超过20000个风力涡轮机

Requirements

要求

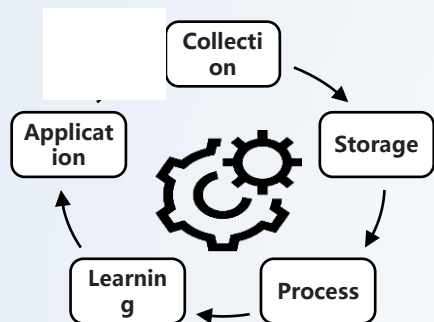
- High speed of ingestion and query/aggregation 读取、访问、整合的高速
- Low cost for storage (historical data) 历史数据的低成本存储
- Integration of query and analytics 访问和分析的整合
- Lightweight 轻体量



- More than 10,000 vehicles 超过10000设备
- Each vehicle has more than 200 sensors (5000+ meters) 每个设备有超过200个传感器 (5000多个仪表)

An Ideal Time Series DB Should Be...

理想的时间序列DB应该是...



Cover the life cycle of data 覆盖数据的生命周期

- Multiple service forms
- 多重服务形式
- Endpoint 终点
- local controller 地方控制器
- data center 数据中心
- Flexible deploy 灵活部署
- Out-Of-the-Box 抛开定势思维
- Support read/write in real time 实时支持读取/编写
- Friendly for analytics 便于分析



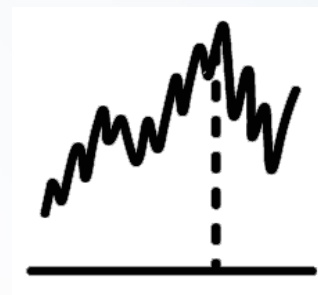
Persist data Efficiently 高效留存数据

- High speed of ingestion 高速读取
- Millions points/sec/node 百万标点/秒/节点
- high compression 大程度压缩
- 1.37bits/point [1] 1.37位/标点
- Lossy and lossless compression 有损压缩和无损压缩
- Millions of time series 上百万个时间序列



Query data with low latency 低延迟查询数据

- Fast Filter data 快速过滤数据
- Hundreds of millisecond on TB data 对TB数据几百毫秒
- Aggregation 整合
- Tens of millisecond on billions of points 对几百亿个标点数万毫秒



Extensive operations of time series [2] 时间序列的延申处理

- Segmentation 分割
- Representation 呈现
- Subsequence matching 子序列匹配
- time-frequency transform 时频转换
- Visualization 可视化



Integration with existing ecosystem 现有生态系统融合

- MatLab
- Spark
- MapReduce
- Grafana
- Kafka

[1] T. Pelkonen *et al.*, "Gorilla: A fast, scalable, in-memory time series database," VLDB, 2015, pp. 1816–1827.

[2] Fu, Tak Chung. "A review on time series data mining." Engineering Applications of Artificial Intelligence 24.1(2011):164-181.

- IoTDB is a time series data management system which supports data acquisition, storage, query and analysis.
- IoTDB是一种支持数据获得、存储、询问和分析的时间序列数据管理系统
- It can be deployed on the endpoint, the local controller, and in the cloud, for single equipment management, monitoring devices in a factory, and mobile equipment service.
- 对于单个设备管理，该系统可以用于端点、本地控制器和云端中，从而监控工厂设备和可移动设备服务



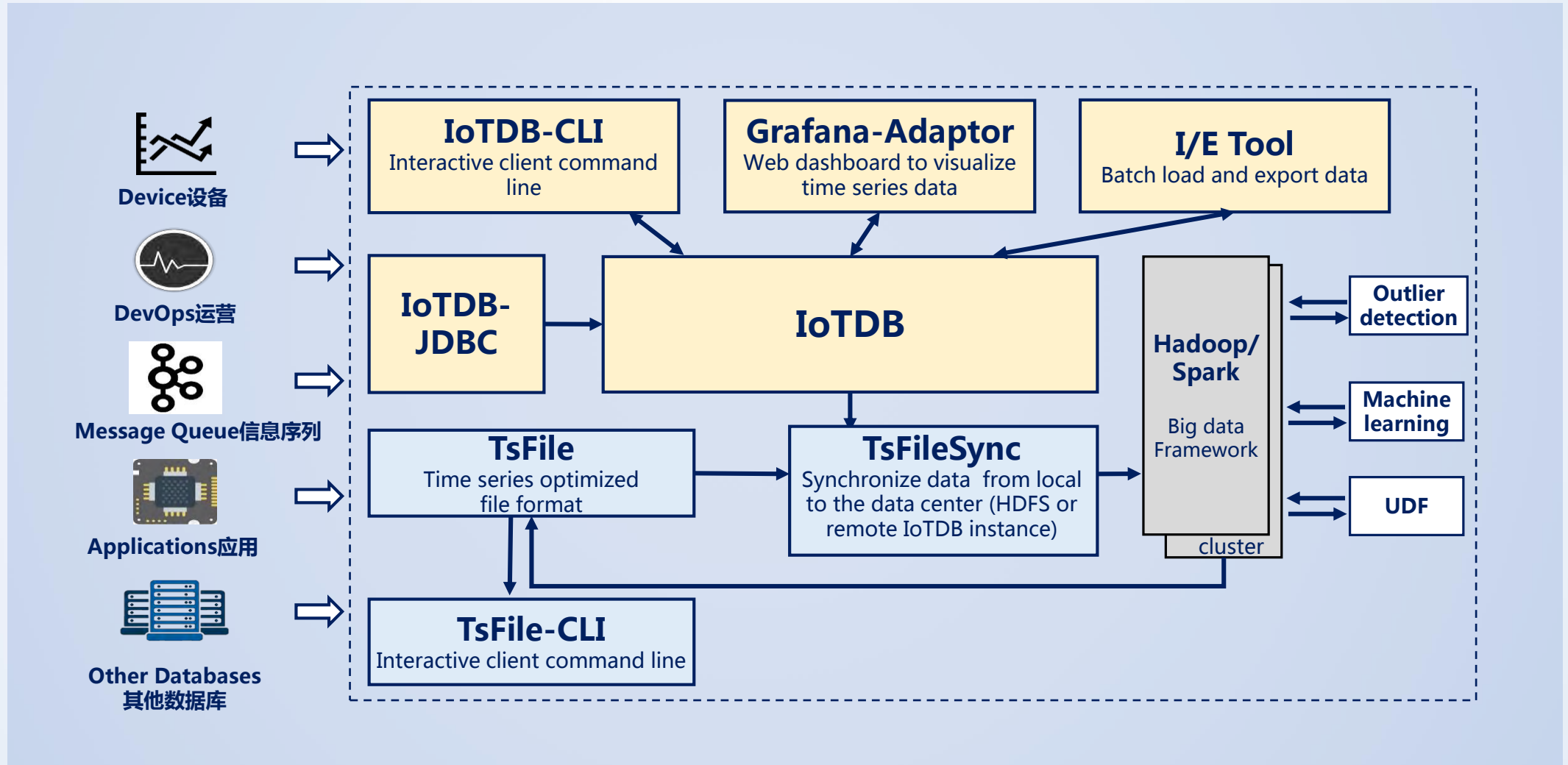
- Manage several time series, support high compression ratio, simple write/query.
- 管理多个时间序列，支持高压缩比和简单读写/访问
- A “zip file” of time series
- 时间序列的压缩包



- Manage multiple time series, support CRUD and advanced query, e.g., aggregation and subsequence matching.
- 管理许多时间序列，支持CRUD和高级访问，如整合和子序列匹配
- A “database” of time series
- 时间序列的数据库



- Manage millions of time series and TBs of data, support distributed analysis framework.
- 管理几百万个时间序列和数据的TB，支持分布式分析框架
- A “data warehouse” of time series
- 时间序列的数据仓库



Industrial Big Data 工业大数据

Physical objects
实体

Domain knowledge & physical model
领域知识和物理模型

Internet of things
Internet of services
物联网 服务互联网

Casual model
因果模型

Object of
Study
研究对象

Base
基础

Data
Source
数据来源

Expectation
期望

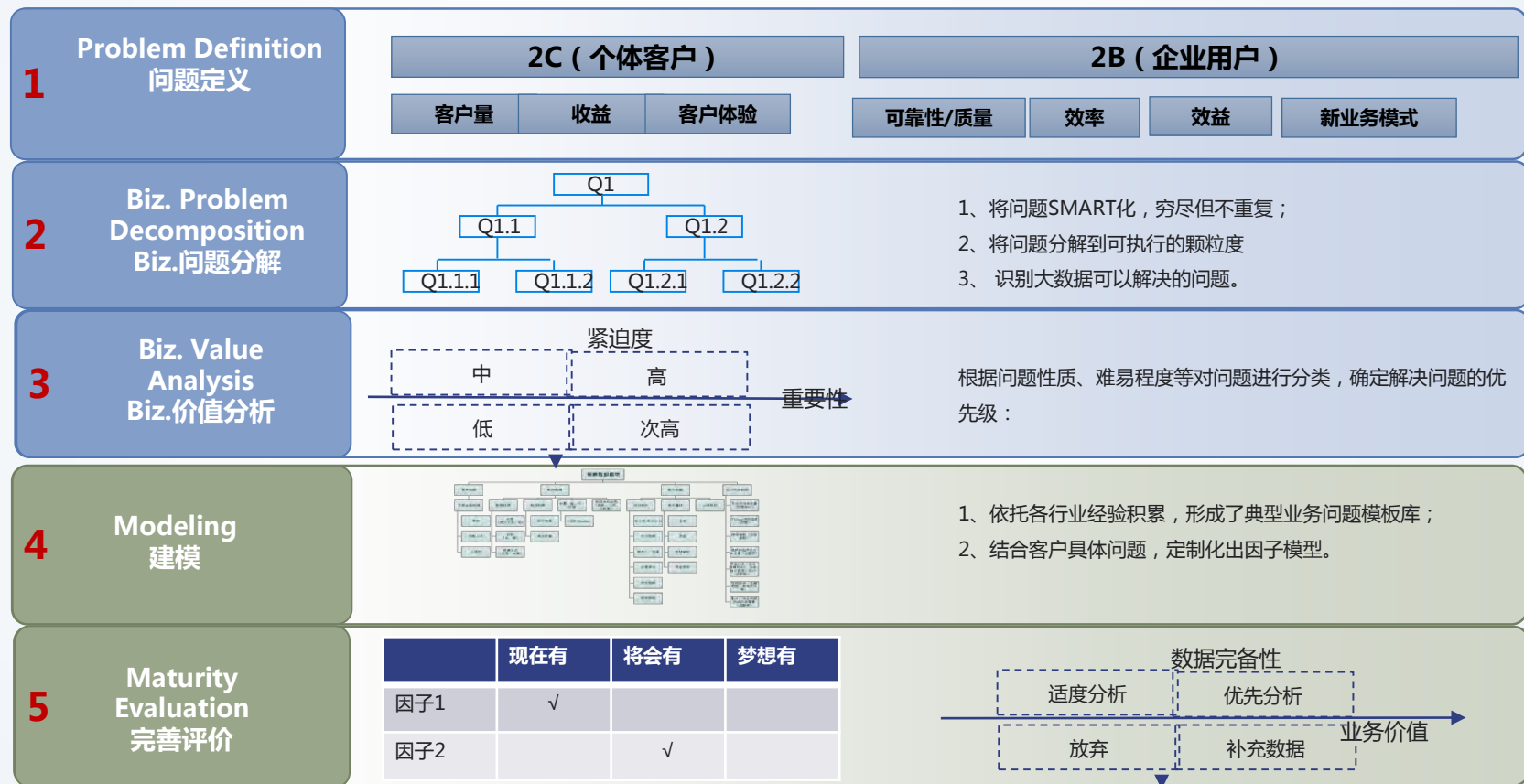
Business Big Data 商业大数据

Internet centric
以互联网为中心

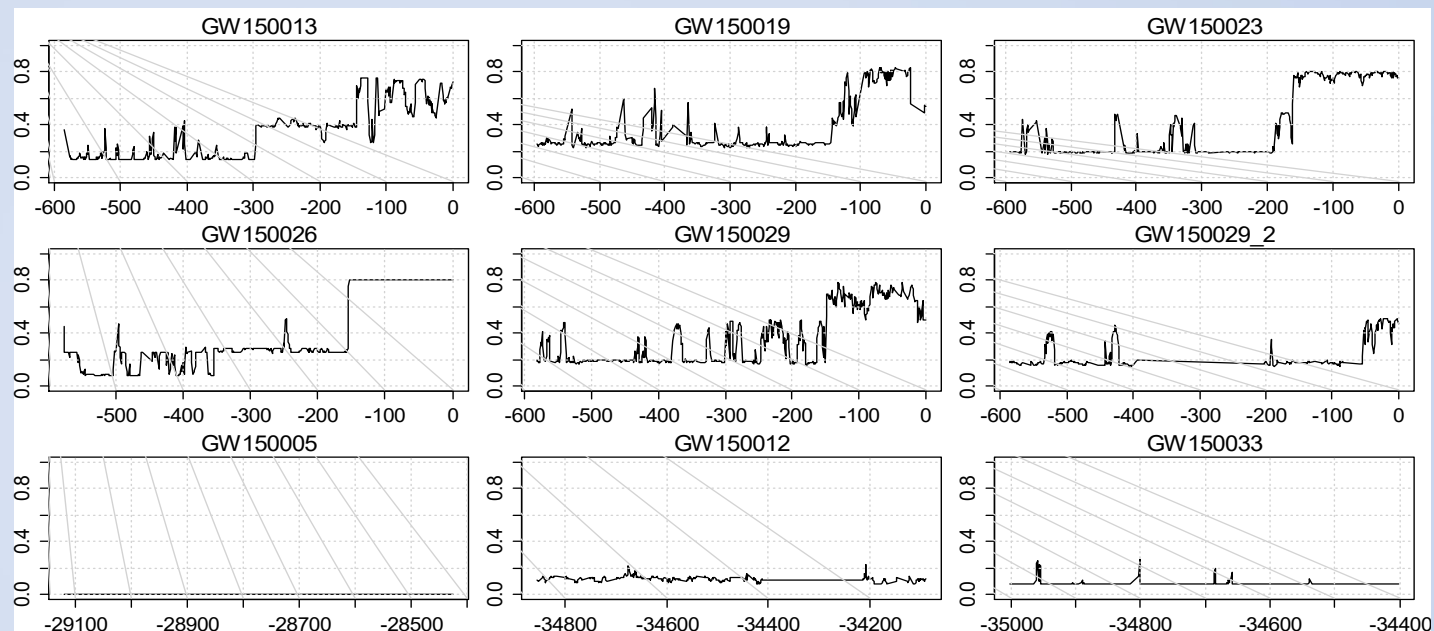
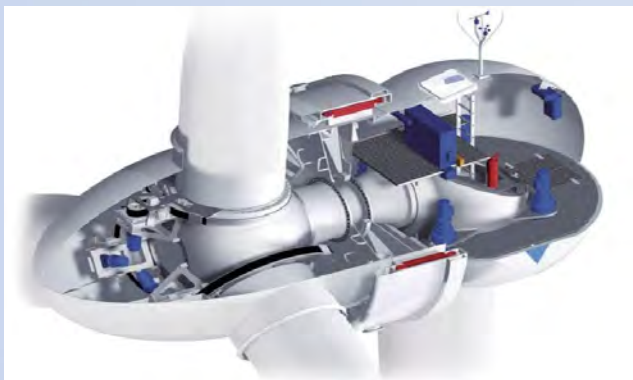
Perception of human being
人类认知

Internet of people
Internet of content
人际网络 内容网络

Statistical correlation
数据更正



Wind Turbine Belt Breakage 风机皮带断裂



Approach 方法

- Noise reduction 降噪
- Decomposition 分解
- Segmentation 分段
- Feature Extraction 特征提取

Intelligent fermentation tank in pharmaceutical factory

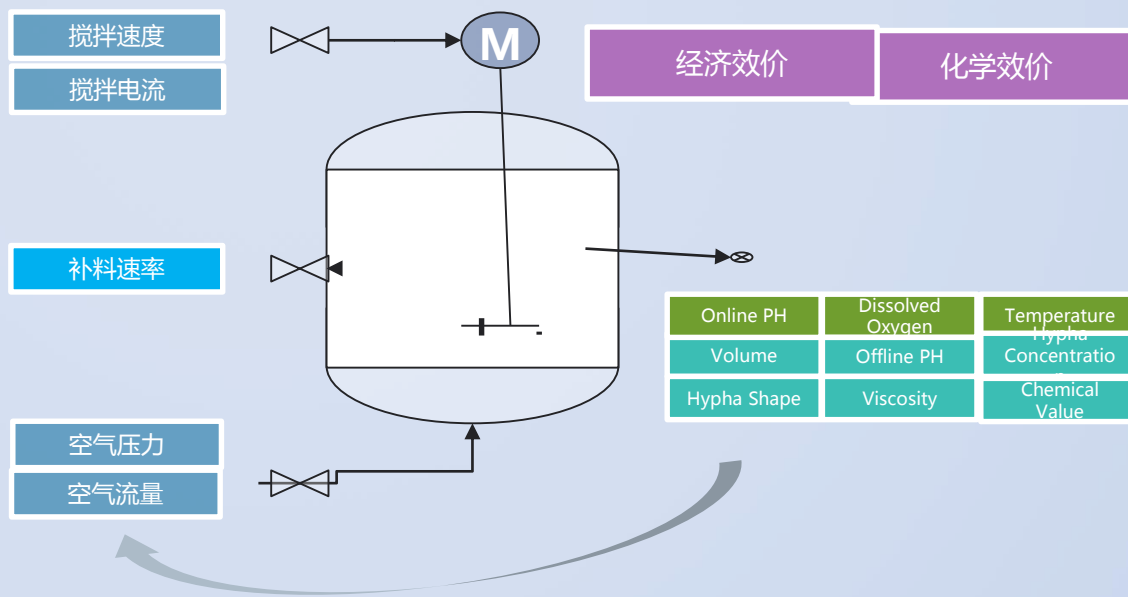
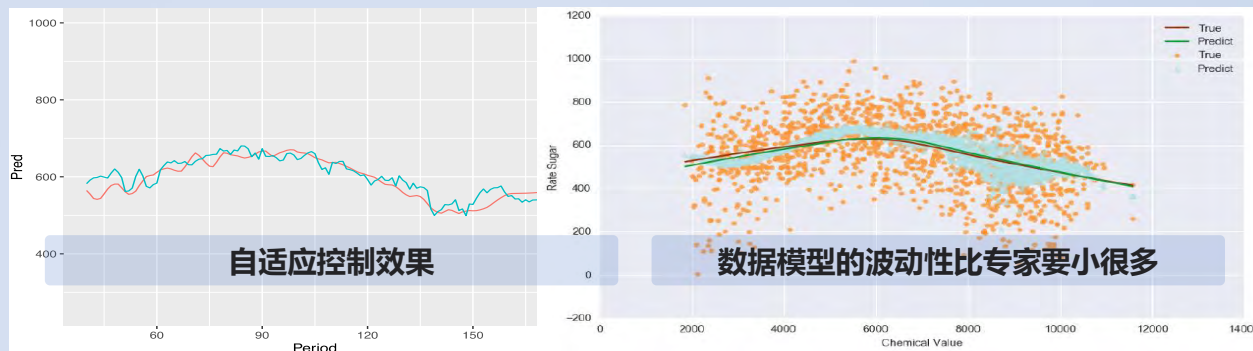
制药厂智能发酵罐

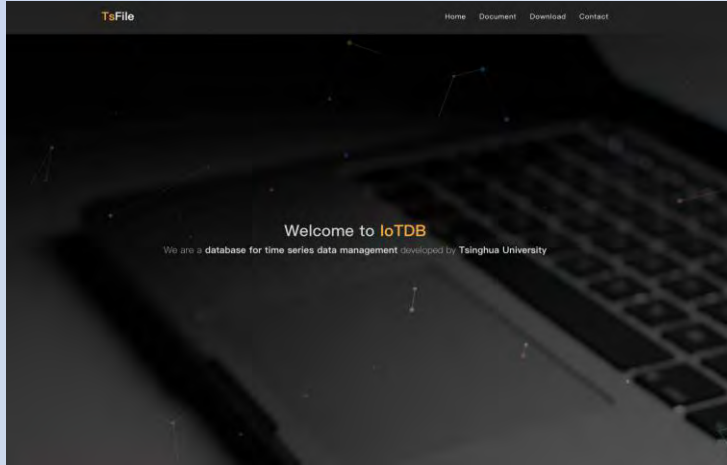
业务目标 智能发酵罐

- 精益控制
- 原料产地分析
- ...

业务问题：发酵过程的精益控制

- 最佳经验控制曲线
- 自适应控制





Binary program

(IoTDB, JDBC, Grafana-Adaptor,
Chinese User Manual)

<http://www.tsfile.org/index>

TsFile

<https://github.com/thulab/tsfile>

IoTDB-JDBC

<https://github.com/thulab/iotdb-jdbc>

SDK

<http://search.maven.org>
cn.edu.tsinghua: tsfile
cn.edu.tsinghua: iotdb-jdbc

Apache Incubator

<https://wiki.apache.org/incubator/IoTDBProposal>



Visit our website - <http://nelbds.org.cn/>

IoTDB, IoTDB-Grafana, IoTDB-Spark, IoTDB-MR are not open sourced
temporary...

融合·协作·共赢

共同把握工业互联网的历史机遇



联盟公众号：工业互联网产业联盟

联盟网址：<http://www.aii-alliance.org/>

联盟邮箱：aii@caict.ac.cn